Example 3. Potato starch grains assessed in size categories

Andrea Onofri, Hans-Peter Piepho and Marcin Kozak

Contents

The dataset

This dataset refers to an experiment which aimed to compare diameters of starch grains from tubers of two potato producers. Starch grains were sampled from tubers collected from the production fields of the producers. The dataset shows the counts of starch grains assigned to one of five diameter classes (< 4, $[4-8[$, $[8-12[$, $[12-16[$, ≥ 16 μm). For each producer, the diameters were measured from twelve photos taken with a microscope.

The original dataset is in a grouped form; one record represents a photo (12 photos per each producer) and shows the number of starch grains in each of the 5 diameter classes. The producer is indicated in the `Group` column. The dataset in this form is available in the `agriCensData` package, which can be installed from gitHub (see here).

```
library(tidyr)
library(nnet)
library(emmeans)
library(survival)
library(agriCensData)

data(starchGrain)
head(starchGrain, 10)
```

```
##    Group Photo c1 c2 c3 c4 c5
## 1     P1     1 21 42 40 29 14
## 2     P1     2 23 45 22 12 10
## 3     P1     3 17 40 33 18  8
## 4     P1     4 24 42 30  8  3
## 5     P1     5 19 16 15  8  3
## 6     P1     6 13  7 11  4 10
## 7     P1     7 34 16  8  5  9
## 8     P1     8 23 13 10  4  1
## 9     P1     9 21 28 26  8  8
## 10    P1    10 26 27 27 14  8
```

An unordered multinomial logit model

The grouped nature of the data suggests we should analyse counts. For each photo, all starch grains should be included in one and only one class, so that the proportions of seeds in the five classes should sum up to one.

Therefore, it's reasonable to consider a multinomial logit model, which treats the classes as nominal categories. In simple terms, a multinomial logit model considers the proportions of grains in each class and how these proportions are affected by the producer. Since such an analysis does not require any assumptions about the real diameter size and its distribution within the population, the method is non-parametric.

In order to fit a multinomial logit model in R, it is useful to reorganise the above dataset into an ungrouped form, so that we have one line for each starch grain (so, 2441 lines), containing all its information: the photo where it was counted, the class to which it was assigned, and the limits of the class (`sizeLow` and `sizeUp`). This reordering can be done using the facilities provided in the `tidyr` package (Wickham and Henry 2018).

```
moltenData <- gather(starchGrain, variable, value, -Group, -Photo)
datasetR <- moltenData[rep(seq_len(nrow(moltenData)),
                           moltenData$value), ][, 1:3]
names(datasetR)[3] <- "Class"
row.names(datasetR) <- 1:2441
rm(moltenData)

#Imputing the diameter interval for each starch grain
datasetR$sizeLow = with(datasetR,
                        ifelse(Class == "c1", NA,
                               ifelse(Class == "c2", 4,
                                      ifelse(Class == "c3", 8,
                                             ifelse(Class == "c4", 12, 16))))
                        )

datasetR$sizeUp = with(datasetR,
                       ifelse(Class == "c1", 4,
                              ifelse(Class == "c2", 8,
                                     ifelse(Class == "c3", 12,
                                            ifelse(Class == "c4", 16, NA))))
                       )

head(datasetR, 10)
```

```
##    Group Photo Class sizeLow sizeUp
## 1     P1     1    c1      NA      4
## 2     P1     1    c1      NA      4
## 3     P1     1    c1      NA      4
## 4     P1     1    c1      NA      4
## 5     P1     1    c1      NA      4
## 6     P1     1    c1      NA      4
## 7     P1     1    c1      NA      4
## 8     P1     1    c1      NA      4
## 9     P1     1    c1      NA      4
## 10    P1     1    c1      NA      4
```

Now we can fit a multinomial model, using the `multinom()` function in the `nnet` package (Venables et al. 2002). The proportions of grains in the five classes for the two producers can be retreived by the multinomial fit with the `emmeans()` function, which we used in the previous examples.

```
mmod <- multinom(Class ~ Group, datasetR)
```

```
## # weights:  15 (8 variable)
## initial  value 3928.637944
## iter  10 value 3802.578106
## final  value 3799.566316
## converged
```

In order to test the overall significance of the producer effect, we can fit a null model (which does not include this effect) and compare it with the previous model, using a likelihood ratio test.

```
mmodNull <- multinom(Class ~ 1, datasetR)
```

```
## # weights:  10 (4 variable)
## initial  value 3928.637944
## final  value 3825.598726
## converged
```

```
anova(mmod, mmodNull)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: Class
##   Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1     1      9760   7651.197
## 2 Group      9756   7599.133 1 vs 2     4 52.06482 1.337066e-10
```

The difference between the producers is significant ($P = 1.34\text{e-}10$). Therefore, we can try to compare the proportions in each class for the two producers. We can do it with the `emmeans()` function, which we can also use to test the difference between the producers (here, for each class).

```
props <- emmeans(mmod, ~ Group | Class)
props
```

```
## Class = c1:
##  Group       prob          SE df   lower.CL  upper.CL
##  P1    0.26387376 0.013294604  8 0.23321635 0.2945312
##  P2    0.26005662 0.011974488  8 0.23244341 0.2876698
##
## Class = c2:
##  Group       prob          SE df   lower.CL  upper.CL
##  P1    0.29117490 0.013704010  8 0.25957340 0.3227764
##  P2    0.21907767 0.011290852  8 0.19304092 0.2451144
##
## Class = c3:
##  Group       prob          SE df   lower.CL  upper.CL
##  P1    0.24112576 0.012903510  8 0.21137022 0.2708813
##  P2    0.20640863 0.011048063  8 0.18093176 0.2318855
##
## Class = c4:
##  Group       prob          SE df   lower.CL  upper.CL
##  P1    0.12102033 0.009838293  8 0.09833319 0.1437075
##  P2    0.15052247 0.009761136  8 0.12801325 0.1730317
##
## Class = c5:
##  Group       prob          SE df   lower.CL  upper.CL
##  P1    0.08280524 0.008313059  8 0.06363529 0.1019752
```

```
##  P2    0.16393459 0.010105997  8 0.14063012 0.1872391
##
## Confidence level used: 0.95
```

```
CLD(props, Letters = letters)
```

```
## Class = c1:
##  Group       prob          SE df   lower.CL  upper.CL .group
##  P2    0.26005662 0.011974488  8 0.23244341 0.2876698  a
##  P1    0.26387376 0.013294604  8 0.23321635 0.2945312  a
##
## Class = c2:
##  Group       prob          SE df   lower.CL  upper.CL .group
##  P2    0.21907767 0.011290852  8 0.19304092 0.2451144  a
##  P1    0.29117490 0.013704010  8 0.25957340 0.3227764   b
##
## Class = c3:
##  Group       prob          SE df   lower.CL  upper.CL .group
##  P2    0.20640863 0.011048063  8 0.18093176 0.2318855  a
##  P1    0.24112576 0.012903510  8 0.21137022 0.2708813  a
##
## Class = c4:
##  Group       prob          SE df   lower.CL  upper.CL .group
##  P1    0.12102033 0.009838293  8 0.09833319 0.1437075  a
##  P2    0.15052247 0.009761136  8 0.12801325 0.1730317  a
##
## Class = c5:
##  Group       prob          SE df   lower.CL  upper.CL .group
##  P1    0.08280524 0.008313059  8 0.06363529 0.1019752  a
##  P2    0.16393459 0.010105997  8 0.14063012 0.1872391   b
##
## Confidence level used: 0.95
## significance level used: alpha = 0.05
```

We see that the producer P1 grew tubers with higher proportions of starch grains in the 5th class and lower in the 2nd.

Although not wrong, this unordered multinomial fit is sub-optimal for several reasons:

1. It is not parsimonious. Indeed, we needed to estimate eight parameters (four proportions per producer, since the fifth one follows from those four estimated). As we will show later, it is possible to fit a meaningful model also with fewer parameters.
2. Indirectly, we demonstrated that the second producer had starch grains of smaller size, although we could not calculate several important statistics, such as the mean diameters for the two producers or the diameter variance. This is because we treated the classes as nominal, neglecting that they had natural ordering and distances.
3. We neglected that starch grains were clustered within photos, meaning that they were stochastically dependent.

A survival model fit

We should recognise that, for each starch grain, we indeed measured the diameter, though with low precision. So, we failed to obtain a value, but instead we obtained an interval. To take into accout the data collection

process, we should model such intervals instead of the corresponding counts.

We can do this by using a survival model, assuming that the distribution of starch grain diameters in the population is Gaussian. As in the other examples, we can use the `survreg` function in the survival package (Therneau 1999).

```
surv.reg1 <- survreg(Surv(sizeLow, sizeUp, type = "interval2") ~ Group,
                     dist = "gaussian", data = datasetR)
summary(surv.reg1)
```

```
##
## Call:
## survreg(formula = Surv(sizeLow, sizeUp, type = "interval2") ~
##     Group, data = datasetR, dist = "gaussian")
##              Value Std. Error     z       p
## (Intercept) 7.3398     0.2122 34.59 < 2e-16
## GroupP2     1.3267     0.2853  4.65 3.3e-06
## Log(scale)  1.8909     0.0206 91.68 < 2e-16
##
## Scale= 6.63
##
## Gaussian distribution
## Loglik(model)= -3824.4   Loglik(intercept only)= -3835.2
##  Chisq= 21.57 on 1 degrees of freedom, p= 3.4e-06
## Number of Newton-Raphson Iterations: 3
## n= 2441
```

We now see that the producer P1 had starch grains with an average diameter of 7.34 μm, and that the difference between P1 and P2 was 1.33 μm, which was significantly different from 0 (P = 3.3e-06).

We argue that the results of this survival model fit are much clearer than those of the multinomial fit, and that, with this second fit, we have better respected the way the data were collected. We discuss these aspects of the analysis in our paper (*Literature reference, when available*).

Do note, however, that our survival model is still not fully OK, because it disregards the point 3 above: we are still neglecting that starch grains are clustered within photos, which means that they are not independent. We will show how to deal with this aspect in the next webpage.

References

Therneau, T (1999) *A Package for Survival Analysis in S*. R package version 2.36-14, Survival

Venables, WN, BD Ripley, WN Venables (2002) *Modern applied statistics with S*4th ed. Springer, New York

Wickham, H, L Henry (2018) *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*